Analysis of Highway-rail Crossing Accidents Regarding Environmental Factors

Bang Ngoc Pham,
School of Engineering and Computer Science,
University of the Pacific

I. Introduction

Traffic accidents are among the most concerning issues in most regions. Of all traffic infrastructure, highway-rail crossing grades are one of the places where many accidents happen. Therefore, researching on factors that can affect the probability of this crash type is considered as an attractive topic for data science. In fact, the causes of highway-rail crossing grade accidents are very various. For my project, I focus on three main environmental factors including temperature, weather condition and visibility in order to determine how accurate it is to predict the driver conditions due to highway-rail crossing grade accidents based on these three factors.

My dataset is obtained from data.gov. It is published by the Federal Railroad Administration - Office of Railroad Safety. It contains 160 variables and more than 243000 observations, which makes it considered as a big dataset. As mentioned above, my project will concentrate on how temperature, weather condition and visibility affects the accident probability, so I mostly used these three variables as the input, and the driver condition as the output.

One of the most challenging when working on this dataset is that all of the necessary variables, except for temperature, are categorical variables while most machine learning algorithms are applied to numeric variables. Since the dataset includes the code columns corresponding to each of these categorical variables, I do not need to encode them. However, it is still not easy to figure out which algorithms can be feasible for these encoded variables. In addition, since the number of uninjured drivers dominates the number of injured and killed drivers, it is challenging to find an approach to predict the driver condition based on the three factors.

II. Related Work

Since traffic accident is one of the common research topics, there has been several works done in this field. In this project, I have taken some reference from three related papers.

The first paper is “Weather impacts on various types of road crashes: a quantitative analysis using generalized additive models” [1], which classifies road crashes into 78 different types and then analyzes how meteorological parameters impact on hourly probabilities of these crash types. In particular, the researchers used additive logistic regression models and found out the non-linear functional relationships between the weather parameters (such as snow, high wind, sun glare, etc.) and the hourly probability of each crash type.

The second paper is “Fatal crashes at highway rail grade crossings: A U.S. based study” [2], which identifies six clusters of risk factors leading to fatal highway-rail crossing crashes in the US within nine years (2010-2018). And one of the clusters is

adverse weather conditions including clear, cloudy and inclement weather. The researchers conducted a descriptive statistics analysis and made some visualizations (including Alluvial plots and plots by TaxicabCA in R) for each cluster. In general, the result shows that for the weather cluster most accidents occur in clear condition. Although this paper does not focus on only weather conditions, it is still a good reference for my project.

The third paper is "A Holistic Analysis of Train-Vehicle Accidents at Highway-Rail Grade Crossings in Florida" [3], in which the researchers investigate a variety of factors that can affect highway-rail grade crossing accidents in Florida between 2010 and 2019, including environmental characteristics. In particular, they used Chi-square test and bar charts to demonstrate the influence of environmental characteristics on the number of accidents. Although this paper is just based on one state, its result can be a reference so that my project can generalize it to the whole country.

III. Solution and Scalability Justification

As mentioned in the previous section, the dataset used in this project is considered large since it contains hundreds of columns and hundreds of thousands of rows, which is approximately 236.2MB. Since this dataset contains hundreds of variables (columns), whenever new observations (rows) are added, it will require a significant amount of memory. Since I am going to use only 4 variables including 3 environmental variables and 1 variable of driver condition, I can quite reduce the amount of necessary memory. However, this data is dynamic because there will be more and more highway-rail crossing grade accidents that may happen in the future and be added into this dataset. Therefore, I need a scalability solution for this problem.

For software scalability, I plan to use Apache Spark to handle large-scale data analytics. Particularly, Apache Spark allows users to perform both exploratory data analysis and machine learning algorithms in a simple, fast, scalable and unified way. Moreover, since my project is implemented using Python, which is also supported by Apache Spark.

For hardware scalability, I can adjust my project in two ways. When the size of this dataset is increased in an appropriate amount, I can scale up my machine vertically first. In detail, I can upgrade my CPUs and memory (perhaps use additional RAM). Nevertheless, vertical scaling is limited to the capacity of only one machine. Thus, when the dataset size is increased by a significant large amount that cannot be handled by one machine, I can scale it out horizontally. In particular, horizontal scaling refers to adding more machines or servers and each server handles a part of the data. In this way, I can compute my data in parallel.

IV. Implementation

1. Programming and Testing Environment

For this project, I save the dataset on my machine with 2 GHz Quad-Core Intel Core i5 and 16GB of memory. For software, I analyze the data using Python on the Jupyter Notebook. In particular, I utilize some libraries in Python to analyze the data: Pandas for data manipulation, Scikit-learn for machine learning algorithms, Matplotlib for visualization and numpy for assisting one of the machine learning algorithms that I apply.

2. Data Extraction

As mentioned above, I focus on three environmental variables: temperature, weather condition, and visibility. In detail, temperature is numeric while the others are all categorical, so I also need the code of those categorical variables, which is also provided in the dataset. Therefore, I need to extract totally 7 variables from the initial dataset. The following table shows the value of each categorical variable and their corresponding codes:

Table 1. Accuracy Scores of 3 Classifier for each Environmental Variables

Code	Weather Condition	Visibility	Driver Condition
1.0	Clear	Dawn	Killed
2.0	Cloudy	Day	Injured
3.0	Rain	Dusk	Uninjured
4.0	Fog	Dark	
5.0	Sleet		
6.0	Snow		

As mentioned above, I utilize the Pandas to extract these three variables along with the driver condition from the original dataset. In particular, I create 6 dataframes and also drop null values in each dataframes:

- 3 dataframes for each environmental code variables and driver condition code
- The other 3 dataframes for these variables and the corresponding number of accidents to each value of the variable by using the groupby() and size() functions in Pandas.

3. Visualization

To gain the general understanding of how temperature, weather conditions and visibility affect the probability of highway-rail crossing grade accidents and driver condition, I conducted three graphs for the three variables vs. the number of accidents colored by driver condition. The library that I use to create the visualization is Matplotlib. SInce temperature is numeric, I plot a stacked histogram for this variable. In contrast, weather condition and visibility are categorical, so I plot two stacked barchart for them.

4. Assess Prediction Accuracy by Machine Learning Algorithms

Another goal of my project is to investigate the accuracy of predicting the driver conditions due to the highway-rail crossing grade accidents. The three environmental factors I focus on are temperature, weather condition, and visibility, together with driver condition, from the original dataset. Since the input and output variables are all categorical (represented by the corresponding codes), except for temperature, I apply classification models. Particularly, for each of these environmental variables vs. driver condition, I implement three algorithms: KNeighbors, Naive Bayes, and Linear Support Vector Machine (Linear SVC) by using Scikit-learn (sklearn) library along with Numpy library.

Unlike KNeigbors and Linear SVC, Naive Bayes is divided into different sub-classifiers: Gaussian, Multinomial, and Bernoulli. For my project, I apply Gaussian for temperature, and Multinomial for both weather condition and visibility.

For each algorithm, I split the data frame corresponding to each variable and driver condition into training sets and test sets, then I fit the model with the training sets to predict the test sets. Then, from those models, I calculate the accuracy score of each variable in predicting the driver condition.

V. Result & Discussion

1. Discussion on Visualization

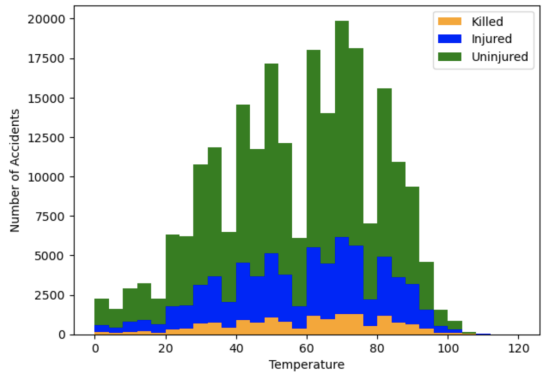

Figure 1. Histogram of Temperature vs. Number of Accidents colored by Driver Condition

Figure 1 shows that accidents are likely to happen when the temperature is between 40 and 75 °F .

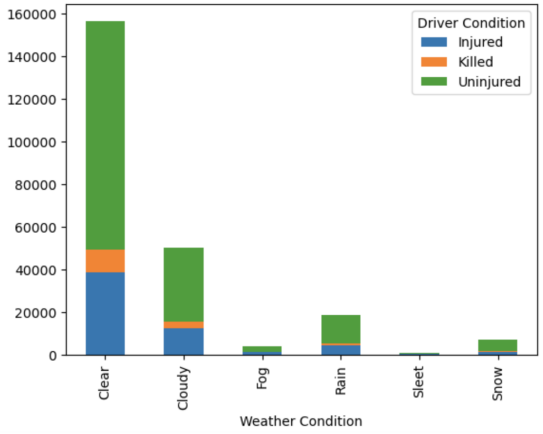

Figure 2. Bar chart of Weather Condition vs. Number of Accidents colored by Driver Condition

Figure 2 shows that in terms of weather conditions, most accidents are under clear, cloudy, rain, snow, fog, sleet in descending order.

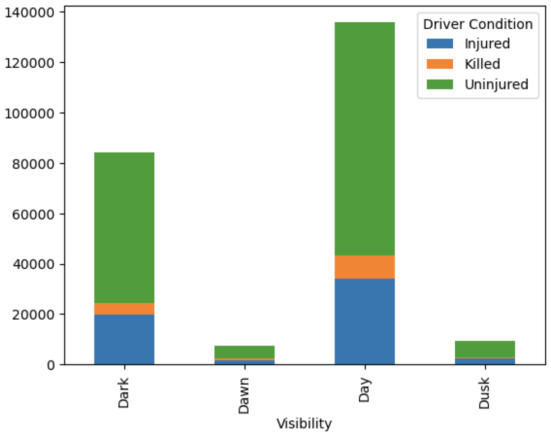

Figure 3. Bar chart of Visibility vs. Number of Accidents colored by Driver Condition

Figure 3 demonstrates that for visibility most accidents are under day, dark, dusk, dawn in descending order.

However, for weather condition and visibility, we also need to consider that people avoid driving in harsh weather conditions, which results in fewer accidents. Hence, we can ignore the domination of clear weather and day visibility to conclude that cloudy and dark can increase the risk of accidents at highway-rail crossing grades.

Based on all of the visualizations, drivers get uninjured in most highway-rail crossing grade accidents under any environmental conditions.

2. Comparison on Accuracy Scores

Table 2. Accuracy Scores of 3 Classifier for each Environmental Variables

	Accuracy Score		
	KNeighbors	Naive Bayes	Linear SVC
Temperature	0.560	0.692	0.692
Weather Condition	0.402	0.692	0.692
Visibility	0.674	0.691	0.692

This result shows that the KNeighbors classifier produces lower accuracy scores for temperature, weather condition and visibility than the other two. On the other hand, Naive Bayes and Linear SVC result in similar accuracy scores for all of the environmental variables. And these scores are nearly 70%, which is moderate-high to suggest that temperature, weather condition and visibility can all be used to predict the driver condition due to the highway-rail crossing grade accidents.

VI. Conclusion

In conclusion, this project has investigated how environmental variables including temperature, weather condition and visibility affects the probability of accidents at highway-rail crossing grades and the driver condition due to the accidents. In particular, the temperature between 40 and 75°F, cloudy weather and dark visibility can increase the risk of highway-rail crossing grade accidents. In most accidents, the drivers get uninjured. Moreover, I also perform three classifiers: KNeighbors, Naive Bayes and LinearSVC on temperature/weather condition/visibility and driver condition. The result shows that both Naive Bayes and LinearSVC are appropriate to predict the driver condition based on each of those environmental variables.

References

[1] Becker, N., Rust, H.W. & Ulbrich, U. (2022). Weather impacts on various types of road crashes: a quantitative analysis using generalized additive models. *Eur. Transp. Res. Rev.* 14, 37. https://doi.org/10.1186/s12544-022-00561-2

[2] Das, S., Kong, X., Lavrenz, M. S., Wu, L., & Jalayer, M. (2022). Fatal crashes at highway rail grade crossings: A U.S. based study. *International Journal of Transportation Science and Technology,* 11(1), 107-117. https://doi.org/10.1016/j.ijtst.2021.03.002.

[3] Singh, P., Pasha, J., Khorram-Manesh, A., Goniewicz, K., Roshani, A., & Dulebenets, M. A. (2021). A Holistic Analysis of Train-Vehicle Accidents at Highway-Rail Grade Crossings in Florida. *Sustainability*, *13*(16), 8842. MDPI AG. http://dx.doi.org/10.3390/su13168842